Internet Engineering Task Force (IETF)
Request for Comments: 6662
Category: Experimental
ISSN: 2070-1721

A. Charny
J. Zhang
Cisco Systems
G. Karagiannis
University of Twente
M. Menth
University of Tuebingen
T. Taylor, Ed.
Huawei Technologies

July 2012

Pre-Congestion Notification (PCN) Boundary-Node Behavior for the Single Marking (SM) Mode of Operation

Abstract

Pre-Congestion Notification (PCN) is a means for protecting the quality of service for inelastic traffic admitted to a Diffserv domain. The overall PCN architecture is described in RFC 5559. This memo is one of a series describing possible boundary-node behaviors for a PCN-domain. The behavior described here is that for a form of measurement-based load control using two PCN marking states: not-marked and excess-traffic-marked. This behavior is known informally as the Single Marking (SM) PCN-boundary-node behavior.

Status of This Memo

This document is not an Internet Standards Track specification; it is published for examination, experimental implementation, and evaluation.

This document defines an Experimental Protocol for the Internet community. This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Not all documents approved by the IESG are a candidate for any level of Internet Standard; see Section 2 of RFC 5741.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at http://www.rfc-editor.org/info/rfc6662.

Table of Contents

1. Introduction

The objective of Pre-Congestion Notification (PCN) is to protect the quality of service (QoS) of inelastic flows within a Diffserv domain, in a simple, scalable, and robust fashion. Two mechanisms are used: admission control to decide whether to admit or block a new flow request and, in abnormal circumstances, flow termination to decide whether to terminate some of the existing flows. To achieve this, the overall rate of PCN-traffic is metered on every link in the PCN-domain, and PCN-packets are appropriately marked when certain configured rates are exceeded. These configured rates are below the rate of the link, thus providing notification to PCN-boundary-nodes about incipient overloads before any congestion occurs (hence the "pre" part of "pre-congestion notification"). The level of marking allows decisions to be made about whether to admit or terminate PCN-flows. For more details, see [RFC5559].

This document describes an experimental edge-node behavior to implement PCN in a network. The experiment may be run in a network in which a substantial proportion of the traffic carried is in the form of inelastic flows and where admission control of micro-flows is applied at the edge. For the effects of PCN to be observable, the committed bandwidth (i.e., level of non-best-effort traffic) on at least some links of the network should be near or at link capacity. The amount of effort required to prepare the network for the experiment (see Section 5.1) may constrain the size of network to which it is applied. The purposes of the experiment are:

- to validate the specification of the SM edge behavior;
- to evaluate the effectiveness of the SM edge behavior in preserving quality of service for admitted flows; and
- to evaluate PCN’s potential for reducing the amount of capital and operational costs in comparison to alternative methods of assuring quality of service.

For the first two objectives, the experiment should run long enough for the network to experience sharp peaks of traffic in at least some directions. It would also be desirable to observe PCN performance in the face of failures in the network. A period on the order of a month or two in busy season may be enough. The third objective is more difficult and could require observation over a period long enough for traffic demand to grow to the point where additional capacity must be provisioned at some points in the network.

Section 3 of this document specifies a detailed set of algorithms and procedures used to implement the PCN mechanisms for the SM mode of operation. Since the algorithms depend on specific metering and marking behavior at the interior nodes, it is also necessary to specify the assumptions made about PCN-interior-node behavior (Section 2). Finally, because PCN uses Diffserv codepoint (DSCP) values to carry its markings, a specification of PCN-boundary-node behavior must include the per-domain behavior (PDB) template specified in [RFC3086], filled out with the appropriate content (Section 4).

Note that the terms "block" or "terminate" actually translate to one or more of several possible courses of action, as discussed in Section 3.6 of [RFC5559]. The choice of which action to take for blocked or terminated flows is a matter of local policy.

A companion document [RFC6661] specifies the Controlled Load (CL) PCN-boundary-node behavior. This document and [RFC6661] have a great deal of text in common. To simplify the task of the reader, the text in the present document that is specific to the SM PCN-boundary-node behavior is preceded by the phrase "[SM-specific]". A similar distinction for CL-specific text is made in [RFC6661].

1.1. Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

This document uses the following terms defined in Section 2 of [RFC5559]:

- PCN-domain
- PCN-ingress-node
- PCN-egress-node
- PCN-interior-node
- PCN-boundary-node
- PCN-flow
- ingress-egress-aggregate
- PCN-excess-rate

- PCN-admissible-rate
- PCN-supportable-rate
- PCN-marked
- excess-traffic-marked

It also uses the terms PCN-traffic and PCN-packet, for which the definition is repeated from [RFC5559] because of their importance to the understanding of the text that follows:

PCN-traffic, PCN-packets, PCN-BA
: A PCN-domain carries traffic of different Diffserv behavior aggregates (BAs) [RFC2474]. The PCN-BA uses the PCN mechanisms to carry PCN-traffic, and the corresponding packets are PCN-packets. The same network will carry traffic of other Diffserv BAs. The PCN-BA is distinguished by a combination of the Diffserv codepoint and the ECN field.

This document uses the following term from [RFC5670]:

- excess-traffic-meter.

To complete the list of borrowed terms, this document reuses the following terms and abbreviations defined in Section 2 of [RFC6660]:

- not-PCN codepoint;
- not-marked (NM) codepoint;
- excess-traffic-marked (ETM) codepoint.

This document defines the following additional terms:

Decision Point
: The node that makes the decision about which flows to admit and to terminate. In a given network deployment, this can be the PCN-ingress-node or a centralized control node. In either case, the PCN-ingress-node is the point where the decisions are enforced.

NM-rate
: The rate of not-marked PCN-traffic received at a PCN-egress-node for a given ingress-egress-aggregate in octets per second. For further details, see Section 3.2.1.

ETM-rate
: The rate of excess-traffic-marked PCN-traffic received at a PCN-egress-node for a given ingress-egress-aggregate in octets per second. For further details, see Section 3.2.1.

PCN-sent-rate
: The rate of PCN-traffic received at a PCN-ingress-node and destined for a given ingress-egress-aggregate in octets per second. For further details, see Section 3.4.

Congestion level estimate (CLE)
: The ratio of PCN-marked to total PCN-traffic (measured in octets) received for a given ingress-egress-aggregate during a given measurement period. The CLE is used to derive the PCN-admission-state (Section 3.3.1) and is also used by the report suppression procedure (Section 3.2.3) if report suppression is activated.

PCN-admission-state
: The state ("admit" or "block") derived by the Decision Point for a given ingress-egress-aggregate based on statistics about PCN-packet marking. The Decision Point decides to admit or block new flows offered to the aggregate based on the current value of the PCN-admission-state. For further details, see Section 3.3.1.

Sustainable aggregate rate (SAR)
: The estimated maximum rate of PCN-traffic that can be carried in a given ingress-egress-aggregate at a given moment without risking degradation of quality of service for the admitted flows. The intention is that if the PCN-sent-rate of every ingress-egress-aggregate passing through a given link is limited to its sustainable aggregate rate, the total rate of PCN-traffic flowing through the link will be limited to the PCN-supportable-rate for that link. An estimate of the sustainable aggregate rate for a given ingress-egress-aggregate is derived as part of the flow termination procedure and is used to determine how much PCN-traffic needs to be terminated. For further details, see Section 3.3.2.

CLE-reporting-threshold
: A configurable value against which the CLE is compared as part of the report suppression procedure. For further details, see Section 3.2.3.

CLE-limit
: A configurable value against which the CLE is compared to determine the PCN-admission-state for a given ingress-egress-aggregate. For further details, see Section 3.3.1.

T_meas
: A configurable time interval that defines the measurement period over which the PCN-egress-node collects statistics relating to PCN-traffic marking. At the end of the interval, the PCN-egress-node calculates the values NM-rate and ETM-rate as defined above and sends a report to the Decision Point, subject to the operation of the report suppression feature. For further details, see Section 3.2.

T_maxsuppress
: A configurable time interval after which the PCN-egress-node MUST send a report to the Decision Point for a given ingress-egress-aggregate regardless of the most recent values of the CLE. This mechanism provides the Decision Point with a periodic confirmation of liveness when report suppression is activated. For further details, see Section 3.2.3.

T_fail
: An interval after which the Decision Point concludes that communication from a given PCN-egress-node has failed if it has received no reports from the PCN-egress-node during that interval. For further details, see Section 3.3.3.

T_crit
: A configurable interval used in the calculation of T_fail. For further details, see Section 3.3.3.

2. [SM-Specific] Assumed Core Network Behavior for SM

This section describes the assumed behavior for PCN-interior-nodes in the PCN-domain. The SM mode of operation assumes that:

- PCN-interior-nodes perform excess-traffic-marking of PCN-packets according to the rules specified in [RFC5670].
- For IP transport, excess-traffic-marking of PCN-packets uses the excess-traffic-marked (ETM) codepoint defined in [RFC6660]; for MPLS transport, an equivalent marking is used as discussed in Appendix C of [RFC6660].
- On each link, the reference rate for the excess-traffic-meter is configured to be equal to the PCN-admissible-rate for the link.
- The set of valid codepoint transitions is as shown in Sections 5.2.1 and 5.2.3.1 of [RFC6660].

3. Node Behaviors

3.1. Overview

This section describes the behavior of the PCN-ingress-node, PCN-egress-node, and the Decision Point (which MAY be collocated with the PCN-ingress-node).

The PCN-egress-node collects the rates of not-marked and excess-traffic-marked PCN-traffic for each ingress-egress-aggregate and reports them to the Decision Point. For a detailed description, see Section 3.2.

The PCN-ingress-node enforces flow admission and termination decisions. It also reports the rate of PCN-traffic sent to a given ingress-egress-aggregate when requested by the Decision Point. For details, see Section 3.4.

Finally, the Decision Point makes flow admission decisions and selects flows to terminate based on the information provided by the PCN-ingress-node and PCN-egress-node for a given ingress-egress-aggregate. For details, see Section 3.3.

Specification of a signaling protocol to report rates to the Decision Point is out of scope of this document. If the PCN-ingress-node is chosen as the Decision Point, [RSVP-PCN] specifies an appropriate signaling protocol.

Section 5.1.2 describes how to derive the filters by means of which PCN-ingress-nodes and PCN-egress-nodes are able to classify incoming packets into ingress-egress-aggregates.

3.2. Behavior of the PCN-Egress-Node

3.2.1. Data Collection

The PCN-egress-node needs to meter the PCN-traffic it receives in order to calculate the following rates for each ingress-egress-aggregate passing through it. These rates SHOULD be calculated at the end of each measurement period based on the PCN-traffic observed during that measurement period. The duration of a measurement period is equal to the configurable value T_meas. For further information, see Section 3.5.

- NM-rate: octets per second of PCN-traffic in PCN-packets that are not-marked (i.e., marked with the NM codepoint);

- ETM-rate: octets per second of PCN-traffic in PCN-packets that are excess-traffic-marked (i.e., marked with the ETM codepoint).

  Note: metering the PCN-traffic continuously and using equal-length measurement intervals minimizes the statistical variance introduced by the measurement process itself. On the other hand, the operation of PCN is not affected if the starting and ending times of the measurement intervals for different ingress-egress-aggregates are different.

3.2.2. Reporting the PCN Data

Unless the report suppression option described in Section 3.2.3 is activated, the PCN-egress-node MUST report the latest values of NM-rate and ETM-rate to the Decision Point each time that it calculates them.

3.2.3. Optional Report Suppression

Report suppression MUST be provided as a configurable option, along with two configurable parameters, the CLE-reporting-threshold and the maximum report suppression interval T_maxsuppress. The default value of the CLE-reporting-threshold is zero. The CLE-reporting-threshold MUST NOT exceed the CLE-limit configured at the Decision Point. For further information on T_maxsuppress, see Section 3.5.

If the report suppression option is enabled, the PCN-egress-node MUST apply the following procedure to decide whether to send a report to the Decision Point, rather than sending a report automatically at the end of each measurement interval.

1. As well as the quantities NM-rate and ETM-rate, the PCN-egress-node MUST calculate the congestion level estimate (CLE) for each measurement interval. The CLE is computed as:

   ```
   [SM-specific]
   CLE = ETM-rate / (NM-rate + ETM-rate)
   ```

   if any PCN-traffic was observed, or CLE = 0 if all the rates are zero.

2. If the CLE calculated for the latest measurement interval is greater than the CLE-reporting-threshold and/or the CLE calculated for the immediately previous interval was greater than the CLE-reporting-threshold, then the PCN-egress-node MUST send a report to the Decision Point. The contents of the report are described below.

The reason for taking into account the CLE of the previous interval is to ensure that the Decision Point gets immediate feedback if the CLE has dropped below the CLE-reporting-threshold. This is essential if the Decision Point is running the flow termination procedure and observing whether (further) flow termination is needed. See Section 3.3.2.

3. If an interval T_maxsuppress has elapsed since the last report was sent to the Decision Point, then the PCN-egress-node MUST send a report to the Decision Point regardless of the CLE value.

4. If neither of the preceding conditions holds, the PCN-egress-node MUST NOT send a report for the latest measurement interval.

Each report sent to the Decision Point when report suppression has been activated MUST contain the values of NM-rate, ETM-rate, and CLE that were calculated for the most recent measurement interval.

The above procedure ensures that at least one report is sent per interval (T_maxsuppress + T_meas). This demonstrates to the Decision Point that both the PCN-egress-node and the communication path between that node and the Decision Point are in operation.

3.3. Behavior at the Decision Point

Operators can choose to use PCN procedures just for flow admission, or just for flow termination, or for both. Decision Points MUST implement both mechanisms, but configurable options MUST be provided to activate or deactivate PCN-based flow admission and flow termination independently of each other at a given Decision Point.

If PCN-based flow termination is enabled but PCN-based flow admission is not, flow termination operates as specified in this document.

Logically, some other system of flow admission control is in operation, but the description of such a system is out of scope of this document and depends on local arrangements.

3.3.1. Flow Admission

The Decision Point determines the PCN-admission-state for a given ingress-egress-aggregate each time it receives a report from the egress node. It makes this determination on the basis of the congestion level estimate (CLE). If the CLE is provided in the egress-node report, the Decision Point SHOULD use the reported value. If the CLE was not provided in the report, the Decision Point MUST calculate it based on the other values provided in the report, using the formula:

[SM-specific]
CLE = ETM-rate / (NM-rate + ETM-rate)

if any PCN-traffic was observed, or CLE = 0 if all the rates are zero.

The Decision Point MUST compare the reported or calculated CLE to a configurable value, the CLE-limit. If the CLE is less than the CLE-limit, the PCN-admission-state for that aggregate MUST be set to "admit"; otherwise, it MUST be set to "block".

If the PCN-admission-state for a given ingress-egress-aggregate is "admit", the Decision Point SHOULD allow new flows to be admitted to that aggregate. If the PCN-admission-state for a given ingress-egress-aggregate is "block", the Decision Point SHOULD NOT allow new flows to be admitted to that aggregate. These actions MAY be modified by policy in specific cases, but such policy intervention risks defeating the purpose of using PCN.

A performance study of this admission control method is presented in [MeLe12].

3.3.2. Flow Termination

[SM-specific] When the PCN-admission-state computed on the basis of the CLE is "block" for the given ingress-egress-aggregate, the Decision Point MUST request the PCN-ingress-node to provide an estimate of the rate (PCN-sent-rate) at which the PCN-ingress-node is receiving PCN-traffic that is destined for the given ingress-egress-aggregate.

If the Decision Point is collocated with the PCN-ingress-node, the request and response are internal operations.

The Decision Point MUST then wait, for both the requested rate from the PCN-ingress-node and the next report from the PCN-egress-node for the ingress-egress-aggregate concerned. If this next egress-node report also includes a non-zero value for the ETM-rate, the Decision Point MUST determine the amount of PCN-traffic to terminate using the following steps:

1. [SM-specific] The sustainable aggregate rate (SAR) for the given ingress-egress-aggregate is estimated using the formula:

   SAR = U * NM-Rate

for the latest reported interval, where U is a configurable factor greater than one and is the same for all ingress-egress-aggregates. In effect, the value of the PCN-supportable-rate for each link is approximated by the expression

U * PCN-admissible-rate

rather than being calculated explicitly.

2. The amount of traffic to be terminated is the difference:

PCN-sent-rate - SAR,

where PCN-sent-rate is the value provided by the PCN-ingress-node.

See Section 3.3.3 for a discussion of appropriate actions if the Decision Point fails to receive a timely response to its request for the PCN-sent-rate.

If the difference calculated in the second step is positive (traffic rate to be terminated), the Decision Point SHOULD select PCN-flows for termination. To that end, the Decision Point MAY use upper rate limits for individual PCN-flows (known, e.g., from resource signaling used to establish the PCN-flows) and select a set of PCN-flows whose sum of upper rate limits is up to the traffic rate to be terminated. Then, these PCN-flows are terminated. The use of upper limits on PCN-flow rates avoids over-termination.

Termination may be continuously needed after consecutive measurement intervals for various reasons, e.g., if the used upper rate limits overestimate the actual flow rates. For such cases it is RECOMMENDED that enough time elapses between successive termination events to allow the effects of previous termination events to be reflected in the measurements upon which the termination decisions are based; otherwise, over-termination may occur. See [Satoh10] and Sections 4.2 and 4.3 of [MeLe10].

In general, the selection of flows for termination MAY be guided by policy.

The Decision Point SHOULD log each round of termination as described in Section 5.2.1.2.

3.3.3. Decision Point Action for Missing PCN-Boundary-Node Reports

The Decision Point SHOULD start a timer t_recvFail when it receives a report from the PCN-egress-node. t_recvFail is reset each time a new report is received from the PCN-egress-node. t_recvFail expires if it reaches the value T_fail. T_fail is calculated according to the following logic:

a. T_fail = the configurable duration T_crit, if report suppression is not deployed;

b. T_fail = T_crit also if report suppression is deployed and the last report received from the PCN-egress-node contained a CLE value greater than CLE-reporting-threshold (Section 3.2.3);

c. T_fail = 3 * T_maxsuppress (Section 3.2.3) if report suppression is deployed and the last report received from the PCN-egress-node contained a CLE value less than or equal to CLE-reporting-threshold.

If timer t_recvFail expires for a given PCN-egress-node, the Decision Point SHOULD notify management. A log format is defined for that purpose in Section 5.2.1.1. Other actions depend on local policy, but MAY include blocking of new flows destined for the PCN-egress-node concerned until another report is received from it. Termination of already admitted flows is also possible, but could be triggered by "Destination unreachable" messages received at the PCN-ingress-node.

If a centralized Decision Point sends a request for the estimated value of PCN-sent-rate to a given PCN-ingress-node and fails to receive a response in a reasonable amount of time, the Decision Point SHOULD repeat the request once. [SM-specific] If the second request to the PCN-ingress-node also fails, the Decision Point SHOULD notify management. The log format defined in Section 5.2.1.1 is also suitable for this case.

> The response timer t_sndFail with upper bound T_crit is specified in Section 3.5. The use of T_crit is an approximation. A more precise limit would be on the order of two round-trip times, plus an allowance for processing at each end, plus an allowance for variance in these values.

See Section 3.5 for suggested values of the configurable durations T_crit and T_maxsuppress.

3.4. Behavior of the Ingress Node

The PCN-ingress-node MUST provide the estimated current rate of PCN-traffic received at that node and destined for a given ingress-egress-aggregate in octets per second (the PCN-sent-rate) when the Decision Point requests it. The way this rate estimate is derived is a matter of implementation.

For example, the rate that the PCN-ingress-node supplies can be based on a quick sample taken at the time the information is required.

3.5. Summary of Timers and Associated Configurable Durations

Here is a summary of the timers used in the procedures just described:

t_meas

Where used: PCN-egress-node.

Used in procedure: data collection (Section 3.2.1).

Incidence: one per ingress-egress-aggregate.

Reset: immediately on expiry.

Expiry: when it reaches the configurable duration T_meas.

Action on expiry: calculate NM-rate and ETM-rate and proceed to the applicable reporting procedure (Section 3.2.2 or Section 3.2.3).

t_maxsuppress

Where used: PCN-egress-node.

Used in procedure: report suppression (Section 3.2.3).

Incidence: one per ingress-egress-aggregate.

Reset: when the next report is sent, either after expiry or because the CLE has exceeded the reporting threshold.

Expiry: when it reaches the configurable duration T_maxsuppress.

Action on expiry: send a report to the Decision Point the next time the reporting procedure (Section 3.2.3) is invoked, regardless of the value of CLE.

t_recvFail

Where used: Decision Point.

Used in procedure: failure detection (Section 3.3.3).

Incidence: one per ingress-egress-aggregate.

Reset: when a report is received for the ingress-egress-aggregate.

Expiry: when it reaches the calculated duration T_fail. As described in Section 3.3.3, T_fail is equal either to the configured duration T_crit or to the calculated value 3 * T_maxsuppress, where T_maxsuppress is a configured duration.

Action on expiry: notify management, and possibly other actions.

t_sndFail

Where used: centralized Decision Point.

Used in procedure: failure detection (Section 3.3.3).

Incidence: only as required, one per outstanding request to a PCN-ingress-node.

Started: when a request for the value of PCN-sent-traffic for a given ingress-egress-aggregate is sent to the PCN-ingress-node.

Terminated without action: when a response is received before expiry.

Expiry: when it reaches the configured duration T_crit.

Action on expiry: as described in Section 3.3.3.

3.5.1. Recommended Values for the Configurable Durations

The timers just described depend on three configurable durations, T_meas, T_maxsuppress, and T_crit. The recommendations given below for the values of these durations are all related to the intended PCN reaction time of 1 to 3 seconds. However, they are based on judgement rather than operational experience or mathematical derivation.

The value of T_meas is RECOMMENDED to be on the order of 100 to 500 ms to provide a reasonable trade-off between demands on network resources (PCN-egress-node and Decision Point processing, network bandwidth) and the time taken to react to impending congestion.

The value of T_maxsuppress is RECOMMENDED to be on the order of 3 to 6 seconds, for similar reasons to those for the choice of T_meas.

The value of T_crit SHOULD NOT be less than 3 * T_meas. Otherwise, it could cause too many management notifications due to transient conditions in the PCN-egress-node or along the signaling path. A reasonable upper bound on T_crit is on the order of 3 seconds.

4. Specification of Diffserv Per-Domain Behavior

This section provides the specification required by [RFC3086] for a per-domain behavior.

4.1. Applicability

This section quotes [RFC5559].

The PCN SM boundary-node behavior specified in this document is applicable to inelastic traffic (particularly video and voice) where quality of service for admitted flows is protected primarily by admission control at the ingress to the domain.

In exceptional circumstances (e.g., due to rerouting as a result of network failures) already admitted flows may be terminated to protect the quality of service of the remaining flows. [SM-specific] The performance results in, e.g., [MeLe10], indicate that the SM boundary node behavior is more likely to terminate too many flows under such circumstances than the CL boundary-node behavior described in [RFC6661].

4.2. Technical Specification

4.2.1. Classification and Traffic Conditioning

Packet classification and treatment at the PCN-ingress-node is described in Section 5.1 of [RFC6660].

PCN packets are further classified as belonging or not belonging to an admitted flow. PCN packets not belonging to an admitted flow are "blocked". (See Section 1 for an understanding of how this term is interpreted.) Packets belonging to an admitted flow are policed to ensure that they adhere to the rate or flowspec that was negotiated during flow admission.

4.2.2. PHB Configuration

The PCN SM boundary-node behavior is a metering and marking behavior rather than a scheduling behavior. As a result, while the encoding uses a single DSCP value, that value can vary from one deployment to another. The PCN working group suggests using admission control for the following service classes (defined in [RFC4594]):

- Telephony (EF)
- Real-time interactive (CS4)
- Broadcast Video (CS3)
- Multimedia Conferencing (AF4)

For a fuller discussion, see Appendix A of [RFC6660].

4.3. Attributes

The purpose of this per-domain behavior is to achieve low loss and jitter for the target class of traffic. The design requirement for PCN was that recovery from overloads through the use of flow termination should happen within 1-3 seconds. PCN probably performs better than that.

4.4. Parameters

The set of parameters that needs to be configured at each PCN-node and at the Decision Point is described in Section 5.1.

4.5. Assumptions

It is assumed that a specific portion of link capacity has been reserved for PCN-traffic.

4.6. Example Uses

The PCN SM behavior may be used to carry real-time traffic, particularly voice and video.

4.7. Environmental Concerns

The PCN SM per-domain behavior could theoretically interfere with the use of end-to-end ECN due to reuse of ECN bits for PCN marking. Section 5.1 of [RFC6660] describes the actions that can be taken to protect ECN signaling. Appendix B of that document provides further discussion of how ECN and PCN can coexist.

4.8. Security Considerations

Please see the security considerations in [RFC5559] as well as those in [RFC2474] and [RFC2475].

5. Operational and Management Considerations

5.1. Deployment of the SM Edge Behavior

Deployment of the PCN Single Marking edge behavior requires the following steps:

- selection of deployment options and global parameter values;
- derivation of per-node and per-link information;
- installation, but not activation, of parameters and policies at all of the nodes in the PCN-domain;
- activation and verification of all behaviors.

5.1.1. Selection of Deployment Options and Global Parameters

The first set of decisions affects the operation of the network as a whole. To begin with, the operator needs to make basic design decisions such as whether the Decision Point is centralized or collocated with the PCN-ingress-nodes, and whether per-flow and aggregate resource signaling as described in [RSVP-PCN] is deployed in the network. After that, the operator needs to decide:

- whether PCN packets will be forwarded unencapsulated or in tunnels between the PCN-ingress-node and the PCN-egress-node. Encapsulation preserves incoming ECN settings and simplifies the PCN-egress-node's job when it comes to relating incoming packets to specific ingress-egress-aggregates, but lowers the path MTU and imposes the extra labor of encapsulation/decapsulation on the PCN-edge-nodes.

- which service classes will be subject to PCN control and what DSCP will be used for each. (See [RFC6660] Appendix A for advice on this topic.)

- the markings to be used at all nodes in the PCN-domain to indicate not-marked (NM) and excess-traffic-marked (ETM) PCN packets;

- the marking rules for re-marking PCN-traffic leaving the PCN-domain;

- whether PCN-based flow admission is enabled;

- whether PCN-based flow termination is enabled.

The following parameters affect the operation of PCN itself. The operator needs to choose:

- the value of CLE-limit if PCN-based flow admission is enabled. [SM-specific] It is RECOMMENDED that the CLE-limit for SM be set fairly low, on the order of 5%.

- the value of the collection interval T_meas. For a recommended range of values, see Section 3.5.1 above.

- whether report suppression is to be enabled at the PCN-egress-nodes and if so, the values of CLE-reporting-threshold and T_maxsuppress. It is reasonable to leave CLE-reporting-threshold at its default value (zero, as specified in Section 3.2.3). For a recommended range of values of T_maxsuppress, see Section 3.5.1 above.

- the value of the duration T_crit, which the Decision Point uses in deciding whether communications with a given PCN-edge-node have failed. For a recommended range of values of T_crit, see Section 3.5.1 above.

- [SM-specific] The factor U that is used in the flow termination procedure (Section 3.3.2). An operational definition for U is given in that section, but it may be thought of as a contingency factor providing a buffer to handle flow peaks above the aggregate

levels expected when flows are admitted. A reasonable value for U is between 1.2 and 2. Larger values of U tend to cause more over-termination of traffic during peaks, but raise the average link utilization level.

5.1.2. Specification of Node- and Link-Specific Parameters

Filters are required at both the PCN-ingress-node and the PCN-egress-node to classify incoming PCN packets by ingress-egress-aggregate. Because of the potential use of multipath routing in domains upstream of the PCN-domain, it is impossible to do such classification reliably at the PCN-egress-node based on the packet header contents as originally received at the PCN-ingress-node. (Packets with the same header contents could enter the PCN-domain at multiple PCN-ingress-nodes.) As a result, the only way to construct such filters reliably is to tunnel the packets from the PCN-ingress-node to the PCN-egress-node.

The PCN-ingress-node needs filters in order to place PCN packets into the right tunnel in the first instance, and also to satisfy requests from the Decision Point for admission rates into specific ingress-egress-aggregates. These filters select the PCN-egress-node, but not necessarily a specific path through the network to that node. As a result, they are likely to be stable even in the face of failures in the network, except when the PCN-egress-node itself becomes unreachable. If all PCN packets will be tunneled, the PCN-ingress-node also needs to know the address of the peer PCN-egress-node associated with each filter.

Operators may wish to give some thought to the provisioning of alternate egress points for some or all ingress-egress-aggregates in case of failure of the PCN-egress-node. This could require the setting up of standby tunnels to these alternate egress points.

Each PCN-egress-node needs filters to classify incoming PCN packets by ingress-egress-aggregate, in order to gather measurements on a per-aggregate basis. If tunneling is used, these filters are constructed on the basis of the identifier of the tunnel from which the incoming packet has emerged (e.g., the source address in the outer header if IP encapsulation is used). The PCN-egress-node also needs to know the address of the Decision Point to which it sends reports for each ingress-egress-aggregate.

A centralized Decision Point needs to have the address of the PCN-ingress-node corresponding to each ingress-egress-aggregate. Security considerations require that information also be prepared for a centralized Decision Point and each PCN-edge-node to allow them to authenticate each other.

Turning to link-specific parameters, the operator needs to derive a value for the PCN-admissible-rate on each link in the network. The first two paragraphs of Section 5.2.2 of [RFC5559] discuss how these values may be derived. ([SM-specific] Confusingly, "PCN-admissible-rate" in the present context corresponds to "PCN-threshold-rate" in the cited paragraphs.)

5.1.3. Installation of Parameters and Policies

As discussed in the previous two sections, every PCN node needs to be provisioned with a number of parameters and policies relating to its behavior in processing incoming packets. The Diffserv MIB [RFC3289] can be useful for this purpose, although it needs to be extended in some cases. This MIB covers packet classification, metering, counting, policing, dropping, and marking. The required extensions specifically include an encapsulation action following reclassification by ingress-egress-aggregate. In addition, the MIB has to be extended to include objects for marking the ECN field in the outer header at the PCN-ingress-node and an extension to the classifiers to include the ECN field at PCN-interior and PCN-egress-nodes. Finally, a new object may need to be defined at the PCN-interior-nodes to represent the packet-size-independent excess-traffic-marking metering algorithm.

The value for the PCN-admissible-rate on each link on a node appears as a metering parameter. Operators should take note of the need to deploy excess-traffic meters either on the ingress or the egress side of each interior link, but not both (Appendix B.2 of [RFC5670].

The following additional information has to be configured by other means (e.g., additional MIBs, NETCONF models).

At the PCN-egress-node:

- the measurement interval T_meas (units of ms, range 50 to 1000);
- whether report suppression is to be applied;
- if so, the interval T_maxsuppress (units of 100 ms, range 1 to 100) and the CLE-reporting-threshold (units of tenths of one percent, range 0 to 1000, default value 0);
- the address of the PCN-ingress-node for each ingress-egress-aggregate, if the Decision Point is collocated with the PCN-ingress-node and [RSVP-PCN] is not deployed;
- the address of the centralized Decision Point to which it sends its reports, if there is one.

At the Decision Point:

- whether PCN-based flow admission is enabled;
- whether PCN-based flow termination is enabled;
- the value of CLE-limit (units of tenths of one percent, range 0 to 1000);
- [SM-specific] the value of the factor U used in the flow termination procedure;
- the value of the interval T_crit (units of 100 ms, range 1 to 100);
- whether report suppression is to be applied;
- if so, the interval T_maxsuppress (units of 100 ms, range 1 to 100) and the CLE-reporting-threshold (units of tenths of one percent, range 0 to 1000, default value 0). These MUST be the same values that are provisioned in the PCN-egress-nodes;
- if the Decision Point is centralized, the address of the PCN-ingress-node (and any other information needed to establish a security association) for each ingress-egress-aggregate.

Depending on the testing strategy, it may be necessary to install the new configuration data in stages. This is discussed further below.

5.1.4. Activation and Verification of All Behaviors

It is certainly not within the scope of this document to advise on testing strategy, which operators undoubtedly have well in hand. Quite possibly an operator will prefer an incremental approach to activation and testing. Implementing the PCN marking scheme at PCN-ingress-nodes, corresponding scheduling behavior in downstream nodes, and re-marking at the PCN-egress-nodes is a large enough step in itself to require thorough testing before going further.

Testing will probably involve the injection of packets at individual nodes and tracking of how the node processes them. This work can make use of the counter capabilities included in the Diffserv MIB. The application of these capabilities to the management of PCN is discussed in the next section.

5.2. Management Considerations

This section focuses on the use of event logging and the use of counters supported by the Diffserv MIB [RFC3289] for the various monitoring tasks involved in management of a PCN network.

5.2.1. Event Logging in the PCN-Domain

It is anticipated that event logging using SYSLOG [RFC5424] will be needed for fault management and potentially for capacity management. Implementations MUST be capable of generating logs for the following events:

- detection of loss of contact between a Decision Point and a PCN-edge-node, as described in Section 3.3.3;
- successful receipt of a report from a PCN-egress-node, following detection of loss of contact with that node;
- flow termination events.

All of these logs are generated by the Decision Point. There is a strong likelihood in the first and third cases that the events are correlated with network failures at a lower level. This has implications for how often specific event types should be reported, so as not to contribute unnecessarily to log buffer overflow. Recommendations on this topic follow for each event report type.

The field names (e.g., HOSTNAME, STRUCTURED-DATA) used in the following subsections are defined in [RFC5424].

5.2.1.1. Logging Loss and Restoration of Contact

Section 3.3.3 describes the circumstances under which the Decision Point may determine that it has lost contact, either with a PCN-ingress-node or a PCN-egress-node, due to failure to receive an expected report. Loss of contact with a PCN-ingress-node is a case primarily applicable when the Decision Point is in a separate node. However, implementations MAY implement logging in the collocated case if the implementation is such that non-response to a request from the Decision Point function can occasionally occur due to processor load or other reasons.

The log reporting the loss of contact with a PCN-ingress-node or PCN-egress-node MUST include the following content:

- The HOSTNAME field MUST identify the Decision Point issuing the log.

- A STRUCTURED-DATA element MUST be present, containing parameters identifying the node for which an expected report has not been received and the type of report lost (ingress or egress). It is RECOMMENDED that the SD-ID for the STRUCTURED-DATA element have the form "PCNNode" (without the quotes), which has been registered with IANA (see [RFC6661] for more information). The node identifier PARAM-NAME is RECOMMENDED to be "ID" (without the quotes). The identifier itself is subject to the preferences expressed in Section 6.2.4 of [RFC5424] for the HOSTNAME field. The report type PARAM-NAME is RECOMMENDED to be "RTyp" (without the quotes). The PARAM-VALUE for the RTyp field MUST be either "ingr" or "egr".

The following values are also RECOMMENDED for the indicated fields in this log, subject to local practice:

- PRI initially set to 115, representing a Facility value of (14) "log alert" and a Severity level of (3) "Error Condition". Note that loss of contact with a PCN-egress-node implies that no new flows will be admitted to one or more ingress-egress-aggregates until contact is restored. The reason a higher severity level (lower value) is not proposed for the initial log is because any corrective action would probably be based on alerts at a lower subsystem level.

- APPNAME set to "PCN" (without the quotes).

- MSGID set to "LOST" (without the quotes).

If contact is not regained with a PCN-egress-node in a reasonable period of time (say, one minute), the log SHOULD be repeated, this time with a PRI value of 113, implying a Facility value of (14) "log alert" and a Severity value of (1) "Alert: action must be taken immediately". The reasoning is that by this time, any more general conditions should have been cleared, and the problem lies specifically with the PCN-egress-node concerned and the PCN application in particular.

Whenever a loss-of-contact log is generated for a PCN-egress-node, a log indicating recovery SHOULD be generated when the Decision Point next receives a report from the node concerned. The log SHOULD have the same content as just described for the loss-of-contact log, with the following differences:

- PRI changes to 117, indicating a Facility value of (14) "log alert" and a Severity of (5) "Notice: normal but significant condition".

- MSGID changes to "RECVD" (without the quotes).

5.2.1.2. Logging Flow Termination Events

Section 3.3.2 describes the process whereby the Decision Point decides that flow termination is required for a given ingress-egress-aggregate, calculates how much flow to terminate, and selects flows for termination. This section describes a log that SHOULD be generated each time such an event occurs. (In the case where termination occurs in multiple rounds, one log SHOULD be generated per round.) The log may be useful in fault management, to indicate the service impact of a fault occurring in a lower-level subsystem. In the absence of network failures, it may also be used as an indication of an urgent need to review capacity utilization along the path of the ingress-egress-aggregate concerned.

The log reporting a flow termination event MUST include the following content:

- The HOSTNAME field MUST identify the Decision Point issuing the log.

- A STRUCTURED-DATA element MUST be present, containing parameters identifying the ingress and egress nodes for the ingress-egress-aggregate concerned, indicating the total amount of flow being terminated, and giving the number of flows terminated to achieve that objective.

  It is RECOMMENDED that the SD-ID for the STRUCTURED-DATA element have the form: "PCNTerm" (without the quotes), which has been registered with IANA (see [RFC6661] for more information). The parameter identifying the ingress node for the ingress-egress-aggregate is RECOMMENDED to have PARAM-NAME "IngrID" (without the quotes). The parameter identifying the egress node for the ingress-egress-aggregate is RECOMMENDED to have PARAM-NAME "EgrID" (without the quotes). Both identifiers are subject to the preferences expressed in Section 6.2.4 of [RFC5424] for the HOSTNAME field.

  The parameter giving the total amount of flow being terminated is RECOMMENDED to have PARAM-NAME "TermRate" (without the quotes). The PARAM-VALUE MUST be the target rate as calculated according to the procedures of Section 3.3.2, as an integer value in thousands of octets per second. The parameter giving the number of flows selected for termination is RECOMMENDED to have PARAM-NAME "FCnt" (without the quotes). The PARAM-VALUE for this parameter MUST be an integer, the number of flows selected.

The following values are also RECOMMENDED for the indicated fields in this log, subject to local practice:

- PRI initially set to 116, representing a Facility value of (14) "log alert" and a Severity level of (4) "Warning: warning conditions".
- APPNAME set to "PCN" (without the quotes).
- MSGID set to "TERM" (without the quotes).

5.2.2. Provision and Use of Counters

The Diffserv MIB [RFC3289] allows for the provision of counters along the various possible processing paths associated with an interface and flow direction. It is RECOMMENDED that the PCN-nodes be instrumented as described below. It is assumed that the cumulative counts so obtained will be collected periodically for use in debugging, fault management, and capacity management.

PCN-ingress-nodes SHOULD provide the following counts for each ingress-egress-aggregate. Since the Diffserv MIB installs counters by interface and direction, aggregation of counts over multiple interfaces may be necessary to obtain total counts by ingress-egress-aggregate. It is expected that such aggregation will be performed by a central system rather than at the PCN-ingress-node.

- total PCN packets and octets that were received for that ingress-egress-aggregate but were dropped;
- total PCN packets and octets admitted to that aggregate.

PCN-interior-nodes SHOULD provide the following counts for each interface, noting that a given packet MUST NOT be counted more than once as it passes through the node:

- total PCN packets and octets dropped;
- total PCN packets and octets forwarded without re-marking;
- total PCN packets and octets re-marked to excess-traffic-marked.

PCN-egress-nodes SHOULD provide the following counts for each ingress-egress-aggregate. As with the PCN-ingress-node, so with the PCN-egress-node it is expected that any necessary aggregation over multiple interfaces will be done by a central system.

- total not-marked PCN packets and octets received;

- total excess-traffic-marked PCN packets and octets received.

The following continuously cumulative counters SHOULD be provided as indicated, but require new MIBs to be defined. If the Decision Point is not collocated with the PCN-ingress-node, the latter SHOULD provide a count of the number of requests for PCN-sent-rate received from the Decision Point and the number of responses returned to the Decision Point. The PCN-egress-node SHOULD provide a count of the number of reports sent to each Decision Point. Each Decision Point SHOULD provide the following:

- total number of requests for PCN-sent-rate sent to each PCN-ingress-node with which it is not collocated;
- total number of reports received from each PCN-egress-node;
- total number of loss-of-contact events detected for each PCN-boundary-node;
- total cumulative duration of "block" state in hundreds of milliseconds for each ingress-egress-aggregate;
- total number of rounds of flow termination exercised for each ingress-egress-aggregate.

6. Security Considerations

[RFC5559] provides a general description of the security considerations for PCN. This memo introduces one new consideration, related to the use of a centralized Decision Point. The Decision Point itself is a trusted entity. However, its use implies the existence of an interface on the PCN-ingress-node through which communication of policy decisions takes place. That interface is a point of vulnerability that must be protected from denial-of-service attacks.

7. Acknowledgements

Ruediger Geib, Philip Eardley, and Bob Briscoe have helped to shape the present document with their comments. Toby Moncaster gave a careful review to get it into shape for Working Group Last Call.

Amongst the authors, Michael Menth deserves special mention for his constant and careful attention to both the technical content of this document and the manner in which it was expressed.

David Harrington’s careful AD review resulted not only in necessary changes throughout the document, but also the addition of the operations and management considerations (Section 5).

Finally, reviews by Joel Halpern and Brian Carpenter helped to clarify how ingress-egress-aggregates are distinguished (Joel) and handling of packets that cannot be carried successfully as PCN-packets (Brian). They also made other suggestions to improve the document, as did Stephen Farrell, Sean Turner, and Pete Resnick.

8. References

8.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC2474] Nichols, K., Blake, S., Baker, F., and D. Black, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers", RFC 2474, December 1998.

[RFC2475] Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., and W. Weiss, "An Architecture for Differentiated Services", RFC 2475, December 1998.

[RFC3086] Nichols, K. and B. Carpenter, "Definition of Differentiated Services Per Domain Behaviors and Rules for their Specification", RFC 3086, April 2001.

[RFC3289] Baker, F., Chan, K., and A. Smith, "Management Information Base for the Differentiated Services Architecture", RFC 3289, May 2002.

[RFC5424] Gerhards, R., "The Syslog Protocol", RFC 5424, March 2009.

[RFC5559] Eardley, P., "Pre-Congestion Notification (PCN) Architecture", RFC 5559, June 2009.

[RFC5670] Eardley, P., "Metering and Marking Behaviour of PCN-Nodes", RFC 5670, November 2009.

[RFC6660] Briscoe, B., Moncaster, T., and M. Menth, "Encoding Three Pre-Congestion Notification (PCN) States in the IP Header Using a Single Diffserv Codepoint (DSCP)", RFC 6660, July 2012.

8.2. Informative References

[MeLe10] Menth, M. and F. Lehrieder, "PCN-Based Measured Rate Termination", Computer Networks Journal (Elsevier), vol. 54, no. 13, pp. 2099-2116, September 2010.

[MeLe12] Menth, M. and F. Lehrieder, "Performance of PCN-Based Admission Control under Challenging Conditions", IEEE/ACM Transactions on Networking, vol. 20, no. 2, April 2012.

[RFC4594] Babiarz, J., Chan, K., and F. Baker, "Configuration Guidelines for DiffServ Service Classes", RFC 4594, August 2006.

[RFC6661] Charny, A., Huang, F., Karagiannis, G., Menth, M., and T. Taylor, Ed., "Pre-Congestion Notification (PCN) Boundary-Node Behavior for the Controlled Load (CL) Mode of Operation", RFC 6661, July 2012.

[RSVP-PCN] Karagiannis, G. and A. Bhargava, "Generic Aggregation of Resource ReSerVation Protocol (RSVP) for IPv4 And IPv6 Reservations over PCN domains", Work in Progress, July 2012.

[Satoh10] Satoh, D. and H. Ueno, "Cause and Countermeasure of Overtermination for PCN-Based Flow Termination", Proceedings of IEEE Symposium on Computers and Communications (ISCC '10), pp. 155-161, Riccione, Italy, June 2010.

Authors' Addresses

Anna Charny
USA

EMail: anna@mwsm.com

Xinyan (Joy) Zhang
Cisco Systems
300 Apollo Drive
Chelmsford, MA 01824
USA

EMail: joyzhang@cisco.com

Georgios Karagiannis
University of Twente
P.O. Box 217
7500 AE Enschede,
The Netherlands

Phone: +31 53 4894099
EMail: g.karagiannis@utwente.nl

Michael Menth
University of Tuebingen
Sand 13
72076 Tuebingen
Germany

Phone: +49-7071-2970505
EMail: menth@uni-tuebingen.de

Tom Taylor (editor)
Huawei Technologies
Ottawa
Canada

EMail: tom.taylor.stds@gmail.com